DOCUMENT RESUME

ED 351 347 TM 019 155

AUTHOR Sireci, Stephen G.
TITLE The Utility of IRT in Small-Sample Testing Applications.
PUB DATE Aug 92
NOTE 22p.; Paper presented at the Annual Meeting of the American Psychological Association (100th, Washington, DC, August 14, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; Financial Services; *Item Response Theory; *Licensing Examinations (Professions); Models; National Programs; *Research Methodology; *Sample Size; *Testing Programs
IDENTIFIERS MULTILOG Computer Program; *One Parameter Model; Three Parameter Model; *Two Parameter Model

ABSTRACT

The utility of modified item response theory (IRT) models in small sample testing applications was studied. The modified IRT models were modifications of the one- and two-parameter logistic models. One-, two-, and three-parameter models were also studied. Test data were from 4 years of a national certification examination for persons desiring certification in personal financial planning. Sample sizes for the 4 years were 173, 149, 106, and 159, respectively. The stability of the item parameters over the 4 years was investigated, and the utility of the models was examined. All IRT analyses used the MULTILOG computer program. Item parameter stability was not exhibited for one-, two-, or three-parameter models or the modifications of the one- and two-parameter logistic models. Other results neither support nor reject the statement that IRT cannot be used with sample sizes smaller than 200 examinees. Five tables and four graphs are included. Two appendixes show MULTILOG input for two models. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
STEPHEN G. SIRECI
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED351347

The Utility of IRT in Small-Sample Testing Applications

Stephen G. Sireci[1]

American Council on Education
GED Testing Service

Presented at the Centennial Annual Convention of the American Psychological Association, Washington, D.C., August 14, 1992

[1]This presentation was supported by a Student Travel Award from the American Psychological Association. The author expresses gratitude for this help and thanks Joan Auchter, Bruce Biskin, David Messersmith, Sen Qi, Lynn Shelley, and Suzette Stone for their assistance in the preparation of this paper. The technical quality of this paper was enhanced through conversations with Maria Potenza, David Thissen, and Howard Wainer. The author thanks these persons for their helpful comments and thanks Dr. Thissen for provision of the MULTILOG and PLOTLOG software programs. The author maintains sole responsibility for the content of this paper and any errors therein. The opinions expressed in this paper are those of the author and do not represent an official position of the American Council on Education.

TM019155

BEST COPY AVAILABLE

The benefits of item response theory (IRT) over classical test theory have been espoused widely by many test specialists (e.g., Hambleton, 1989; Lord, 1980). These test specialists assert that IRT offers test developers increased measurement precision, and so tests developed using IRT provide accurate assessment of examinee ability (or proficiency) using fewer items than tests developed using classical procedures.

The efficient precision of measurement provided by IRT is accomplished by placing person and item parameters on the same measurement scale (i.e., item and person parameters are scaled in the metric of the underlying latent trait). Because the person and item parameters are on the same measurement scale, they are *sample-independent*. That is, the person parameters (ability estimates) are independent of the particular sample of items administered, and the item parameters (difficulty, discrimination, guessability) are independent of the particular sample of examinees tested. This feature of IRT allows for direct equating of tests assembled from a common pool of items, and provides an unambiguous means for combining information provided by different item types onto a common scale.

Though IRT offers many benefits to test developers, it has one clear limitation: relatively large numbers of examinees (sample sizes) must be tested to provide accurate results. This limitation is unfortunate because many tests are administered to, and developed from, relatively small numbers of examinees. For this reason, most applications of IRT in test use and development are found in large-scale testing organizations.

Previous research on IRT with small samples has concluded that sample sizes under 200 are not appropriate for even the simplest (i.e., least general) IRT models (e.g., one-parameter logistic model) and that much larger samples are required for the more complex (e.g., two- and three-parameter) models (c.f. Hulin, Lissak, & Drasgow, 1982; Lord, 1968; Ree & Jensen, 1980; Thissen & Wainer, 1982; Wright & Stone, 1979). However, some recent research investigating modifications of these traditional IRT models has indicated that modified IRT models may be appropriate for use in some small-scale testing applications (Barnes & Wise, 1991; Sireci, 1991).

This study investigated the utility of modified IRT models in a small-sample testing application. The modified IRT models used were modifications of the one- and two-parameter logistic models. The purpose

of this investigation was to determine whether these modified models would be appropriate in small-sample testing applications.

Data

The test data analyzed in this study were part of a national certification examination for persons desiring certification in personal financial planning. The data represented four separate administrations of the examination over a four-year period. Because the requirements to sit for the examination were fairly stringent, only about 150 persons sat for the examination each year. The number of examinees (sample sizes) who sat for the examination each year was 173, 149, 106, and 159, for years 1 through 4, respectively. The examination was comprised of 100 multiple-choice items, and separate test forms were administered each year. The test forms were constructed to be parallel and were equated using a common-item (nonequivalent groups) linear equating procedure (Angoff, 1984; Kolen & Brennan, 1987). There were 13 items in common among the four test forms. The data for these 13 items were aggregated over the four-year period so that comparisons could be made between the small-sample data (i.e., the data from a single test administration) and the aggregate data (i.e., the data combined for the 13 items over the four-year period).

Item Parameter Stability

The first part of the investigation evaluated the stability of the item parameters over the four-year period. Item parameter stability was evaluated by using restricted and unrestricted IRT models and comparing their fit to the data. The unrestricted IRT models computed the item parameters for each group separately, while the restricted models constrained the item parameters to be equal among the four groups. Thus, the restricted models represented item parameter stability (item parameters were equal from sample to sample), and the unrestricted models represented item parameter instability (i.e., the item parameters were not equal across samples).[2]

---

[2]Restricted IRT models have been used previously in a variety of research contexts. For example, Stone and Lane (1991) used restricted IRT models to investigate item parameter stability over time; Thissen, Steinberg, and Gerrard (1986), and Thissen, Steinberg, and Wainer (1988, in press) used restricted IRT models to investigate differential item functioning; and Wainer, Sireci, and Thissen (1991) used restricted IRT models to investigate differential testlet functioning.

The purpose of the analysis of item parameter stability was to determine whether an IRT model could be directly applied to a small-sample data set. If item parameter stability was exhibited over the four groups, then the item parameters would be appropriate for estimating examinee proficiency. Using restricted IRT models, Sireci (1991) found that item parameter stability did not hold over three separate small-sample test administrations. However, the IRT models in the Sireci (1991) study did not include a fixed lower-asymptote, which Barnes and Wise (1991) suggested for use with small data sets.

Mixed IRT Models

The second part of the present investigation evaluated the utility of "mixed" IRT models for small data sets. Mixed IRT models use more than one IRT model in a single analysis. Using a mixed IRT model, some test items could be modeled using a 1PL, while other items could be modeled using a 2PL, and etc. Thissen (1991) demonstrated how mixed IRT models can be used to include different item types (e.g., multiple-choice items and categorical items) in a single analysis run. The purpose of using mixed IRT models in the present study was to demonstrate how incorporation of prior information (i.e., incorporation of item parameters based on an aggregated data set) can increase the precision of IRT estimates based on small samples.

The One-, Two-, and Three-Parameter IRT Models

The three IRT models used in this study were the one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL). There are several thorough descriptions of these and other IRT models available in the literature (e.g., Hambleton, 1989; Lord & Novick, 1968; and Thissen & Steinberg, 1986), and so they are not described in detail here. The equations for the 1PL, 2PL, and 3PL, respectively are presented below:

$$p(\Theta) = \frac{1}{1+\exp[-\bar{a}(\Theta - b)]} \quad (1)$$

$$p(\Theta) = \frac{1}{1+\exp[-a(\Theta - b)]} \quad (2)$$

$$p(\Theta) = \frac{c+(1-c)}{1+\exp[-a(\Theta - b)]} \quad (3)$$

where P (Θ) is the probability of choosing the correct answer as a function of Θ; b is the difficulty level of the item, a is the slope of the item characteristic curve (ICC) at the point $\Theta = b$, and c is the lower asymptote of the ICC. The item parameters a, b, and c are commonly referred to as the discrimination, difficulty, and lower-asymptote (or guessing) parameters, respectively. [$\bar{a}$ is fixed in the 1PL and indicates a constant value of discrimination.] The restricted IRT models used in this study involved constraining the a, b, and/or c parameters to be equal for identical items taken by examinees in one of the four different groups. The modified IRT models used in this study involved fixing one or more of these parameters to be equal to some pre-specified value.

Comparing Model Fit: -2logs and X^2

All IRT analyses reported here were conducted using the MULTILOG (version 6.0) IRT software program (Thissen, 1991). MULTILOG is a very general program that fits a variety of IRT models to test data using the marginal maximum likelihood method (Bock & Aitken, 1981). MULTILOG uses a maximum likelihood procedure and so "negative twice the log likelihood" values (-2loglks) are provided for each analysis. Because the difference between the -2loglks of two competing (i.e., hierarchical) IRT models is distributed as chi-square, this difference can be evaluated for statistical significance by computing the probability of obtaining the observed difference by chance (with degrees of freedom equal to the difference between the number of free parameters estimated in each model). If the additional parameters in the more general (unrestricted) model adds substantially to the data-model fit, then the difference between the -2loglks will be significant. However, if the difference is not significant, then the more parsimonious (i.e., restricted) model is preferred. This chi-square difference test is appropriate only for comparing hierarchical models (i.e., the more general model estimates all of the parameters of the restricted model, plus some additional ones).

Procedure and Results

Assessing dimensionality. To determine whether the test items were appropriate for IRT analysis, an inter-item tetrachoric correlation matrix was computed for the 13-item data set based on the aggregated data of 587 (173+149+106+159) examinees. A one-dimensional (factor) model was fit to this inter-item correlation matrix using LISREL-7 (Joreskog & Sorbom, 1988). This preliminary analysis was conducted to determine whether the unidimensional assumption of IRT was satisfied.

The one-factor model accounted for over 66% of the variance in the data and exhibited low values of residual error (coefficient of determination was .66, RMSE=.06, and the standard errors for the items ranged from .049 to .051). Although the assessment of test dimensionality using item intercorrelations is controversial (cf. Gorsuch, 1983; Green, 1983) these results were considered to be indicative of unidimensionality, and so the IRT models were deemed appropriate.

Determining model fit for the aggregated data. The 1PL, 2PL, and 3PL models were fit to the aggregated data set to determine the most appropriate model for these data. Priors for the lower asymptotes (*c* parameters) were set at .25, which was the reciprocal of the number of response alternatives. The results of these analyses are presented in Table 1. The significance tests of the differences between the -2loglks of the three models indicated that the 2PL was the appropriate model for these data. This improvement in fit of the 2PL over the 1PL is consistent with a preliminary analysis of the data that indicated moderate variation among the item biserials (thus undermining the constant discrimination assumption of the 1PL). The lack of improvement in fit for the 3PL may represent either the absence of a guessing factor among the less-proficient examinees, or may result from an inability to compute accurate lower asymptotes because of the relatively small sample size (Thissen & Wainer, 1982).

Table 1
Results of 1PL, 2PL and 3PL Analyses on Aggregated Data (N=587)

Model	-2loglk	# Free Parameters	Difference X^2	*df*	p
3PL	1578	39			
2PL	1587	26	9	13	.78
1PL	1627	14	49	25	.002

Because of the reported difficulty in estimating lower asymptotes (*cj's*) from relatively small data sets, and because Barnes & Wise (1991) recommended incorporating a fixed value for the asymptotes into a one-parameter model, modified 1PL (MOD-1PL) and modified 2PL (MOD-2PL) analyses were conducted. These modified models added a fixed constant lower asymptote to the 1PL and 2PL. The cj's for both modified

models were fixed at .20, which was the reciprocal of the number of response alternatives minus .05. This value was used based on previous research by Barnes and Wise (1991) and Divgi (1984). The results for the MOD-1PL and MOD-2PL analyses are presented in Table 2. The MOD-1PL analysis resulted in a smaller -2loglk than the 1PL; however, this loglikelihood was significantly different from that obtained by the MOD-2PL. Therefore, it appears that the assumption of equal slopes (discrimination) among the items is not appropriate for these data. The -2loglk for the MOD-2PL was identical to the value obtained in the 2PL analysis and so it is unclear whether the addition of the lower asymptote improves the performance of the 2PL model.

Table 2
Results of MOD-1PL and MOD-2PL Analyses on Aggregated Data ("MOD" indicates the inclusion of fixed, non-zero cj's) (N=587)

Model	-2loglk	# Free Parameters	Difference X^2	*df*	p
MOD-2PL	1587	26	9	13	.78
MOD-1PL	1627	14	49	25	.002

Determining item parameter stability. To determine whether any of the IRT models could be applied directly to a small-sample data set (i.e., the data from one of the four groups), the stability of the item parameters across the four groups (samples) was investigated.[3] The results of these analyses are reported in Table 3. [The input command file used to fit the MOD-2PL model is presented in Appendix A to illustrate how the constraints were imposed via MULTILOG.] A comparison of the restricted (item parameter stability) and unrestricted (instability) -2loglks indicated that item parameter stability was not exhibited for the 1PL, 2PL or 3PL models. Therefore, direct application of these IRT models to any one of the samples would not be appropriate.

[3]Differences in proficiency between the examinee samples was not expected to affect the results of this analysis. Analysis of the mean theta values for each group were not statistically significant. Furthermore, Stone and Lane (1991) reported that item parameter stability held over time for groups that differed in proficiency.

Aside from the traditional 1PL, 2PL, and 3PL models, the stability of the item parameters resulting from four other IRT models was investigated (also reported in Table 3). The first two models represent restricted 2PL models that were investigated by Sireci (1991). The model labeled"*aj's* only" constrained only the slopes of the 2PL to be equal among the four samples. The model "*bj's* only" restricted only the location (difficulty) parameters to be equal among the four groups. Though these two models exhibited better fit than the fully-restricted 2PL, neither the *aj's* nor *bj's* exhibited satisfactory stability. The other two models investigated were the modified 1PL (MOD-1PL) and 2PL (MOD-2PL) models that incorporated a fixed value of .20 for the lower asymptote parameters. These models also failed to exhibit stability over the four small-sample data sets and so it was concluded that none of the IRT models studied were appropriate for these small-sample data.

Table 3
Results of Item Parameter Stability Analyses

Model	-2loglk	# Free Parameters	Difference X^2	*df*	p
3PL					
Unrestricted	1302	159			
Restricted	1548	42	246	117	<.001
2PL					
Unrestricted	1305	107			
Restricted	1557	29	252	78	<.001
aj's only	1364	68	59	39	.018
bj's only	1372	68	67	39	.003
1PL					
Unrestricted	1409	56			
Restricted	1610	17	201	39	<.001
MOD-2PL					
Unrestricted	1306	107			
Restricted	1548	29	242	78	<.001
MOD-1PL					
Unrestricted	1397	56			
Restricted	1560	17	163	39	<.001

The Utility of Mixed IRT Models and Aggregated Data

The preceding analyses indicated that the IRT models used were not appropriate for the small sample data. However, it is possible that the item parameters obtained from the aggregated data analyses are appropriate (i.e., stable). Though there is no way to determine the stability of the parameters estimated from the aggregated data (aside from waiting several years to cross-validate on a new aggregated sample), we can investigate whether the item parameters obtained from the aggregated data are beneficial in calibration of parameters in a single (small) sample run. The purpose of this section is to determine whether such aggregated data can be beneficial to the small-sample test practitioner.

If common items exist across several small-sample administrations of a test (as was the case with the present study), then the data on these common items could be aggregated over administrations. The item parameters obtained from analysis of the aggregated data are likely to be more stable than those based on the small-sample administrations. If appropriate, these item parameters could then be used for item selection, scaling, equating, and scoring of subsequent test forms.

IRT item parameters based on aggregated data. To investigate the utility of using aggregated data, some item parameters from the 2PL analysis based on the aggregated data were selected for inclusion in a mixed-model analysis on the data for a single administration (Group 4, n=159). The item parameters that resulted from the 2PL analysis on the aggregated data set are presented along with the content area specification for each item (for the five content areas measured by this test) in Table 4 . Although a few parameters have high standard errors, these standard errors are very small in relation to the standard errors observed in the unrestricted models reported above (i.e., based on separate calibrations for each sample). The data in Table 4 represent typical data that can be computed readily by the small-sample test practitioner who has several items in common over separate administrations of an examination. Because many small-sample test forms are equated using common-item equating procedures, it is likely that many of these practitioners could easily create such aggregated data sets. The five items that were selected for the mixed-model IRT analysis (MIX) are highlighted in Table 4.

Table 4

Item Parameters from 2PL Analyses on Aggregated Data (N=587)

Item	Content Area	*aj* (s.e.)	*bj* (s.e.)
1	A	.33 (.11)	1.72 (.66)
2	A	.10 (.10)	-2.94 (3.14)
3	A	.82 (.15)	-1.61 (.30)
4	B	.68 (.15)	-2.58 (.53)
5	B	.24 (.11)	.66 (.55)
6	C	.98 (.15)	- .53 (.12)
7	C	1.22 (.17)	- .93 (.12)
8	C	.51 (.15)	-3.67 (.99)
9	C	.53 (.13)	-1.97 (.49)
10	D	.60 (.12)	- .15 (.18)
11	D	.41 (.11)	-1.15 (.39)
12	D	.56 (.12)	- .90 (.26)
13	E	.62 (.14)	-2.24 (.47)

Note: Values are scaled to Mu=0.0 and SD=1.0
Items in **boldface** indicate items selected for mixed analysis reported below.

In selecting items to be used on future test forms, both statistical and content criteria must be satisfied. Therefore, a resourceful test developer would most likely select items within each content area that fulfill the content specifications of the test and demonstrate satisfactory statistical criteria. Such statistical criteria would include satisfactory difficulty and discrimination values. Furthermore, in testing situations where cut-off scores are used, such as in licensure or certification testing, the test developer would also want to select items that maximize discrimination (test information) at the cut-score.

Given such considerations, the test developer could select items based on the aggregated IRT parameter estimates and incorporate these estimates into a mixed model analysis. Items could be selected that maximize the information around the cut-score (given necessary content constraints). The parameters for the re-used (common) items could be

fixed at their values obtained from the aggregated data, while the new items on a test form could be fit using a parsimonious IRT model such as the 1PL or MOD-1PL. Using this procedure, relatively few item parameters would need to be estimated and the need for large amounts of data would be obviated.

To test whether this procedure would result in increased test information for these data, one item from each of the five content areas was selected based upon the 2PL difficulty and discrimination parameters and their standard errors (the selected item numbers are printed in **boldface** in Table 4). A mixed-model IRT analysis (MIX) was performed on the data from one of the samples (Group 4, n=159), and 1PL and MOD-1PL analyses were applied to the same data. An additional model, MOD-MIX was also applied. This model added a fixed lower asymptote (at .20) to the MIX model . The input command file for the MOD-MIX analysis is reproduced in the Appendix B. The commands in this input file illustrate how to impose the necessary equality constraints among the new (1PL) items, and how to fix the *cj* parameters for all 13 items, and the *aj* and *bj* parameters for the 5 common items.

To evaluate the relative contribution of the prior information (i.e., the 2PL item parameters estimated from the aggregated data) item and test characteristic curves were computed for four IRT models that were fit to the data. The test information curves were computed for each of the models to determine whether increased test information was obtained by using the parameters based on the aggregated data.

Test information. Test information curves (TIC) depict the reciprocal of the standard error values at any point along the ability scale. Thus, larger amounts of information indicate smaller amounts of measurement error. The preferred shape of a TIC varies according to the purpose of the test (Lord, 1977; Hambleton, 1989; Thissen, 1990). For tests that are designed to discriminate between examinees along the entire continuum of proficiency (theta), platykrtic (flat) curves are preferable. For tests that use cut-off scores, leptokurtic curves are preferred that peak (maximize information) at the level of theta that corresponds to the cut-score (and so skewness would be determined by the location of the cut-score). Regardless of the shape desired, tests that generate TICs that have larger upper asymptotes are preferable to those with lower upper asymptotes.

Figure 1 presents the test information curve (TIC) resulting from a 1PL analysis of the Group 4 data. Figure 2 presents the TIC for the

MOD-1PL analysis for these same data. A comparison of Figures 1 and 2 reveals that inclusion of the fixed *cj's* increased test information along the theta (Θ) range -1 to +3. However, the 1PL exhibited greater information at the lower end of the Θ-scale.

Figure 1: *Test Information Curve for 1PL (Group 4 Data)*

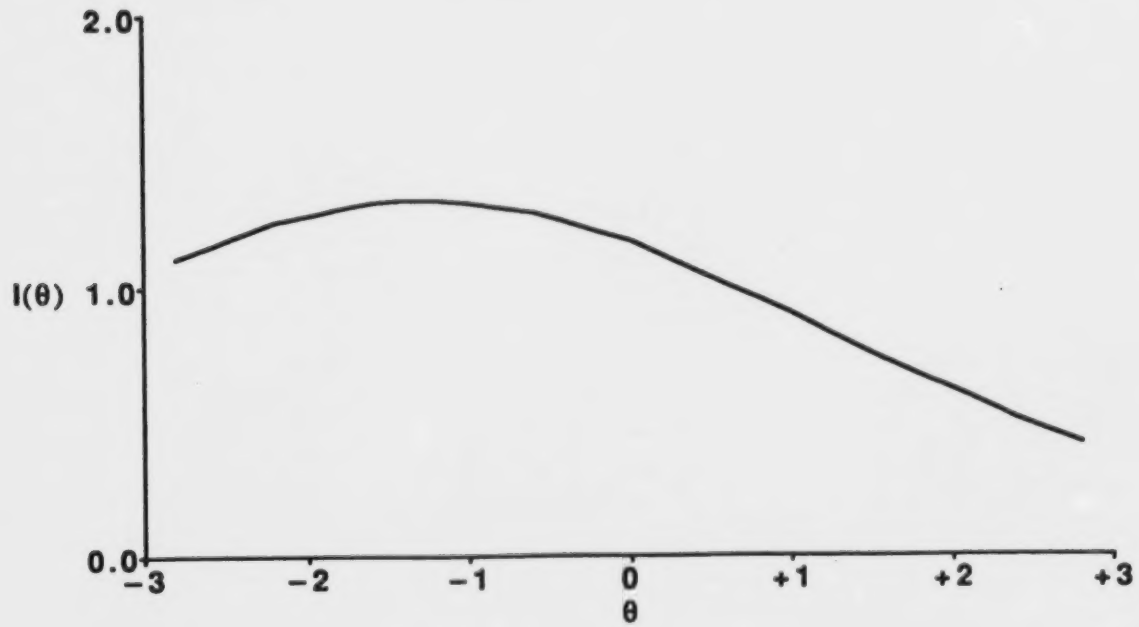


Figure 2: *Test Information Curve for MOD-1PL (includes fixed cj's)*

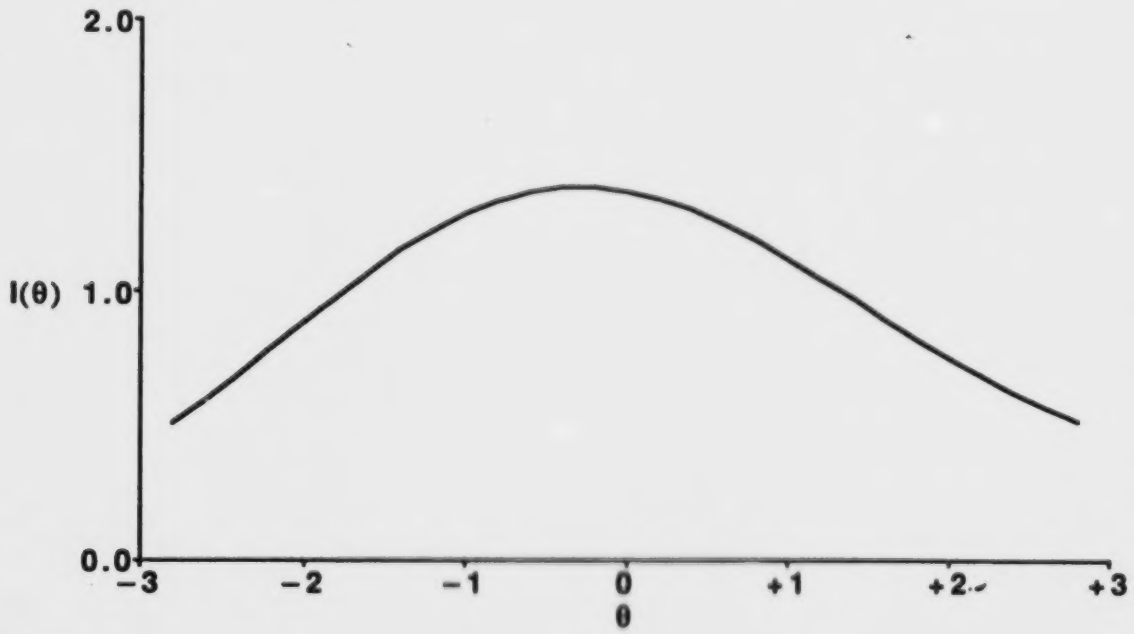


Figure 3 presents the TIC for the MIX analysis. The incorporation of the fixed parameter values from the aggregated 2PL analysis did not increase test information at any point along the Θ-scale. However, when the fixed *cj's* were incorporated into the model (MOD-MIX), the shape of the TIC changed dramatically. The TIC for the MOD-MIX model is illustrated in Figure 4. The incorporation of the fixed *cj's* increased the test information substantially over the Θ-scale range -2 to +.5, and appears to peak at Θ=-1. This value of theta is equivalent to one standard deviation unit below the population mean and is a common cut-off score used by many licensing and certification programs. Thus, the TIC produced by the MOD-MIX analysis may be useful in these application areas.

Figure 3: *Test Information Curve for MIX Model (Group 4 data)*

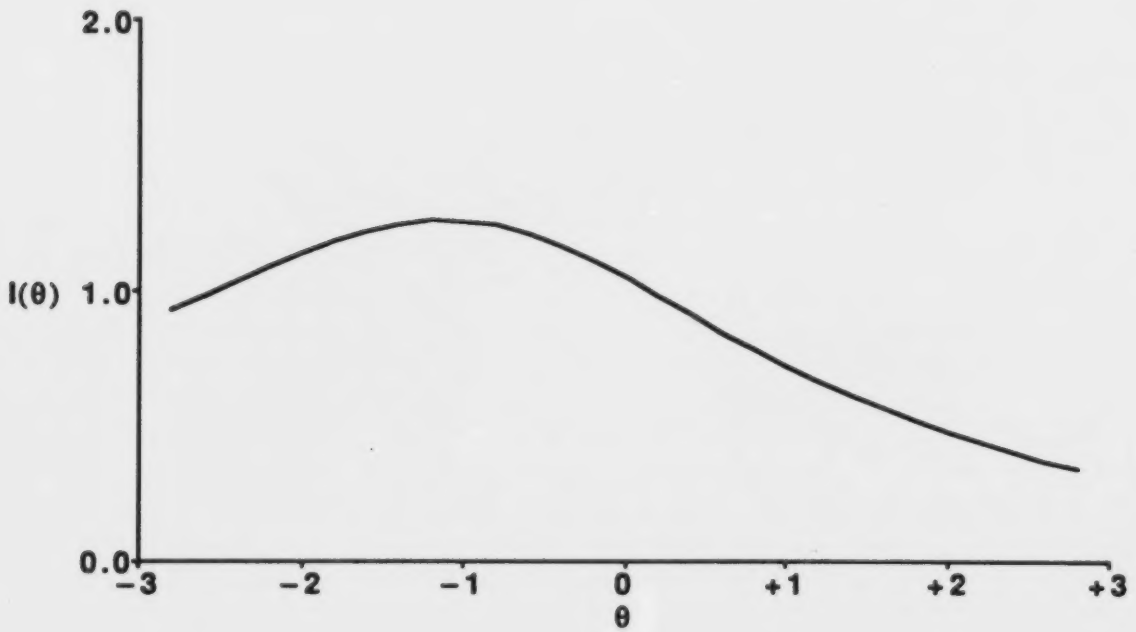
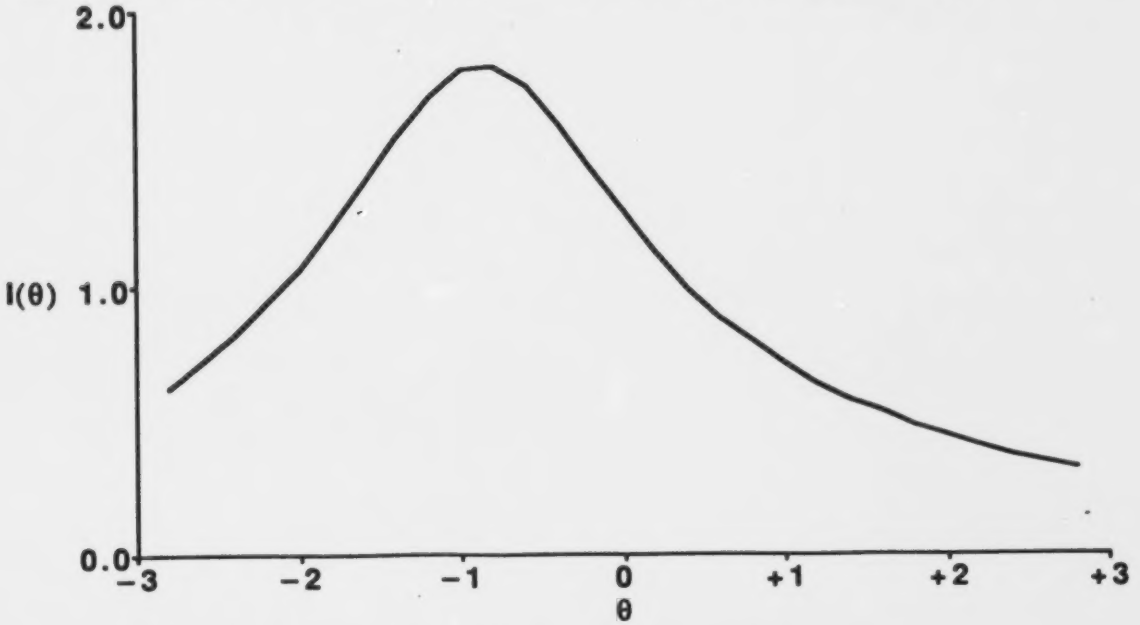


Figure 4: *Test Information Curve for MOD-MIX (includes fixed cj's)*



RMSE. Though the test information curves are informative regarding measurement precision, the standard errors associated with the score of each examinee provides another means for evaluating the precision of IRT-based scores. To evaluate the precision of the MOD-MIX model, root mean square residual errors (RMSE) were calculated for the group 4 examinees using both the 1PL and MOD-MIX models. The RMSE index is used widely in simulation studies to estimate the degree of departure of IRT estimates from their known parameters (e.g., Barnes & Wise, 1991; Thissen, 1990). Though the "true" proficiency estimates of the examinees in the study were not known, their proficiency estimates provided by the aggregate analysis can serve as a reference for the proficiency estimates computed from the single-sample runs. Thus, the RMSE were computed by taking the difference between each examinee's ability estimate ("theta-hat") from the MOD-2PL analysis (using the aggregated data) and an alternative model (either MOD-MIX or 1PL), and then taking the square root of the average of these squared differences.

Table 5 presents the mean, standard deviation, and range of the standard errors of the theta-estimates, for the 159 examinees in Group 4, for each of the three models of interest. The RMSE for the MOD-MIX and 1PL models (using the MOD-2PL as the "true" model) are also provided. The MOD-MIX model exhibited the smallest average standard error, although it also exhibited a higher RMSE than did the 1PL model.

The differences observed between the standard errors and RMSE are difficult to interpret; particularly in light of the relatively small variation among the estimates for the 1PL. Further analyses were planned to replicate the RMSE analyses with the other three sample, but unfortunately, these analyses could not be completed in time for this presentation. However, it is likely that the RMSE analyses may not be appropriate because the true theta values are not known.

Table 5
RMSE and Standard Errors of 1PL and MOD-MIX Models

Model	Avg. S.E.	St. Dev.	Range	RMSE
MOD-2PL	.69	.03	.658 - .796	
MOD-MIX	.66	.05	.597 - .783	.0471
1PL	.69	.02	.657 - .755	.0261

Discussion

This investigation has first, demonstrated that restricted modeling can be used to investigate item parameter stability over small-samples of real test data, and second, investigated a means by which some small-sample test practitioners may benefit from IRT methodology. Though the problems with using IRT under small-sample conditions have been noted since the early days of IRT (Lord, 1968), little research has been done to redress this problem. Perhaps the bottom line is that IRT cannot be used with sample sizes smaller than 200 examinees, no matter how much we incorporate prior information and/or fiddle with the parameter estimation procedure. The results of this study neither reject nor support such a statement, and so it is clear that future research is need in this area. Though the challenge is great, the effort will be justified if IRT can be brought into the hands of small-sample test practitioners.

One avenue for future research may be to increase the number of items for which aggregated data are available and include them in a calibration run for an actual examination. In this study, only five items incorporated prior information (aside from the fixed lower asymptotes on all items in the MOD models), and they were analyzed together with only eight other items. The contribution of prior information to longer test

lengths, and the inclusion of prior information on more of the test items, are likely to improve test information. Future research should also focus on selecting items that will maximize a target test information curve.

A rather atypical feature of this study was that fairly complex IRT models were applied to these data, yet relatively small numbers of parameters were estimated. For example, in the MOD-MIX model, essentially a 3PL model was fit to the data, yet only 8 parameters were estimated for the 13 items! This reduction in the number of parameters to be estimated stems from the fixing of the *cj's* for all items, and the fixing of the *aj's* and *bj's* for the five "common" items. Because the data/parameter ratio is the keystone for robust parameter estimation, any promise for the use of IRT with small data sets must concentrate on increasing that ratio. Though fixing item parameters reduces the number of parameters to be estimated by the model, it invokes the critical question "How defensible are the parameter values that are fixed in these runs?" The research of Divgi (1984), Barnes & Wise (1991) suggests that fixed *cj's* are defensible; however their findings must be replicated with real test data. Though this study offers promise for IRT application in small-sample settings, the stability of the item parameters gathered from test data aggregated over several small-sample administrations requires further investigation.

References

Angoff, W.H. (1971). Scales, Norms, and Equivalent Scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd edt.). Washington, D.C.: American Council on Education, pp. 508-600. [Reprinted by Educational Testing Service, Princeton, NJ, 1984.]

Barnes, L.B., & Wise, S.L. (1991). The utility of a modified one-parameter IRT model with small sample sizes. *Applied Measurement in Education, 4,* 143-157.

Bock, R.D., & Aitken (1981). Marginal maximum likelihood estimation of item parameters: aplication of an algorithm. *Psychometrika, 46,* 443-459.

Davison, M.L., (1985). Multidimensional scaling versus components analysis of test intercorrelations. *Psychological Bulletin, 97,* 94-105.

Divgi, D.R., (1984). Does small N justify use of the Rasch model? Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Green, S.B., (1983). Identifiability of spurious factors with linear factor analysis with binary items. *Applied Psychological Measurement, 7,* 139-147.

Gorsuch, R.L., (1983). *Factor Analysis.* Hillsdale, N.J.: Lawrence Erlbaum Associates,

Hambleton, (1989). Principals and selected appilications of item response theory. In R.L. Linn (Ed.), *Educational Measurement,* (3rd edt.), Washington, D.C.: American Council on Education, pp. 147-200.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: a Monte Carlo study. *Applied Psychological Measurement, 6,* 249-260.

Joreskog, K.G., & Sorbom, D. (1988) *LISREL 7 User's Reference Guide.* Mooresville, IN: Scientific Sortware.

Kolen, M.J., & Brennan, R.L., (1987). Linear equating models for the common item nonequivalent populations design. *Applied Psychological*

Measurement, 11, 263-277.

Lord, F.M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989-1020.

Lord, F.M., (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117-138.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Ree, M.J., & Jensen, H.E., (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. Weiss (ed.), *Proceedings of the 1979 Computerized Adaptive Testing Conference.* Minneapolis: University of Minnesota.

Sireci, S.G., (1991). "Sample-independent" item parameters? An investigation of the stability of IRT item parameters estimated from small data sets. Paper presented at the annual conference of the Northeastern Educational Research Association, Ellenville, New York, October, 1991. [Available through ERIC Clearinghouse, document number ED338707.]

Stone, C.A., & Lane, S. (1991). Use of restricted item response theory models for examining the stability of item parameter estimates over time. *Applied Measurement in Education, 4,* 125-141.

Thissen, D. (!990). Reliability and measurement precision. In H. Wainer (Ed.) *Computerized adaptive testing: a primer.* Hillsdale, N.J.: In H. Wainer (Ed.), *Computerized adaptive testing: a primer.* Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 161-186.

Thissen, D. (1991). *MULTILOG (Version 6.0) User's Guide.* Mooresville, IN: Scientific Sortware.

Thissen & Steinberg, L. 91986). A taxonomy or item response models. *Psychometrika, 49,* 501-519.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychological Bulletin, 99,* 118-128.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H.Wainer & H. Braun (Eds.), *Test Validity*. Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 147-169.

Thissen, D., Steinberg, L., & Wainer, H. (1992). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397-412.

Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: a primer*. Hillsdale, N.J.: Lawrence Erlbaum Associates, pp. 65-102.

Wainer, H., Sireci, S.G., & Thissen, D (1991). Differential testlet functioning: definition and detection. *Journal of Educational Measurement, 28*, 197-219.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339-368.

Wright, B.D., 7 Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Zwick, R. (1991). Effects of item-order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*, 10-15.

Appendix A

MULTILOG INPUT FOR MOD-2PL MODEL
(i.e., restricted (stability) model with fixed lower asymptotes):

```
>PRO RA IN NI=52 NG=4 NE=587;

>TEST ALL L3;

>EQUAL AJ IT=(14(1)26) WI=(1(1)13);
>EQUAL AJ IT=(27(1)39) WI=(14(1)26);
>EQUAL AJ IT=(40(1)52) WI=(27(1)39);
>EQUAL BJ IT=(14(1)26) WI=(1(1)13);
>EQUAL BJ IT=(14(1)26) WI=(1(1)13);
>EQUAL BJ IT=(14(1)26) WI=(1(1)13);

>FIX ALL CJ VA=.20;

>END;
```

APPENDIX B

MULTILOG INPUT FOR MOD-MIX MODEL: (Fixed parameters for items 3,5,7,10,&13 based on aggregated 2PL; fixed lower asymptotes).

```
>PRO RA IN NI=13 NG=1 NE=159;
>TEST ALL L3;

>EQUAL AJ IT=(8,9,11,12) WI=(1,2,4,6);
>EQUAL AJ IT=2 WI=1;
>EQUAL AJ IT=4 WI=1;
>EQUAL AJ IT=6 WI=4;
>EQUAL AJ IT=9 WI=8;
>EQUAL AJ IT=11 WI=9;
>EQUAL AJ IT=12 WI=11;

>FIX ALL CJ VA=.20;

>FIX IT=3 AJ VA=.82;
>FIX IT=3 BJ VA=-1.61;
>FIX IT=5 AJ VA=.24;
>FIX IT=5 BJ VA=.66;
>FIX IT=7 AJ VA=1.22;
>FIX IT=7 BJ VA=-.93;
>FIX IT=10 AJ VA=.60;
>FIX IT=10 BJ VA=-.15;
>FIX IT=13 AJ VA=.62;
>FIX IT=13 BJ VA=-2.24;

>END;
```